Dataset:
https://drive.google.com/drive/folders/1ndKQHxWHjUiXwlZR9-Hg0NNVHRAJAN2b?usp=sharing

Dataset: Stock price and volume traded information for salient stocks at minute-level over a 3 year period from 2017 to 2020.

Problem Statement:

The fraud control unit at a major regulator would like to probe historical stock trades for abnormal stock trading behaviours. To start with, any of the below events that happen should be considered an anomaly:

- A1: Current trade for a particular stock is at a price that deviates from previous minute's trade close price by more than 0.5% (over or under).
- A2: Traded volume in a particular stock is more than 2% above the average traded volume for the last 10 minutes prior.

Whenever an anomaly is found, then the trade that triggered the anomaly should be captured for further scrutiny.

Your task is to identify all anomalies, emit the trades that are anomalous along with the type of the anomaly that got detected. Note that if a trade is anomalous due to any one definition, it suffices to emit that trade without needing to check that trade for other anomalies.

Suggestions:

1. First, convert the batch data into a data stream by loading into Kafka using Spark batch.
2. When doing so, pay attention to the quality of the data.
3. Second, write your spark streaming code pointing to that Kafka topic such that stream is read minute by minute and the anomaly testing happens using helper constructs like window functions. While testing your code, you may want to use a smaller time window for streaming (e.g. 10 seconds) so that you are not waiting for the next stream window.

Notes:

- Demonstrate live outputs
- Do not assume the data files are sorted in ascending order by time
- Do not assume the data has no gaps in dates for each stock & date combination.

- Assume that there will be bad rows

Please Note:

Don't discuss anything with others and whatever you know please do it within the allotted time and submit the code + screenshots of the process and screen recording of the outputs.

Submission should be uploaded to folder :

Final_Assignment_<date of submission>_<FirstName>